H. M. Keerthi Kumar* and B. S. Harish

A New Feature Selection Method for Sentiment Analysis in Short Text

https://doi.org/10.1515/jisys-2018-0171
Received March 19, 2018.

Abstract: In recent internet era, micro-blogging sites produce enormous amount of short textual information, which appears in the form of opinions or sentiments of users. Sentiment analysis is a challenging task in short text, due to use of formal language, misspellings, and shortened forms of words, which leads to high dimensionality and sparsity. In order to deal with these challenges, this paper proposes a novel, simple, and yet effective feature selection method, to select frequently distributed features related to each class. In this paper, the feature selection method is based on class-wise information, to identify the relevant feature related to each class. We evaluate the proposed feature selection method by comparing with existing feature selection methods like chi-square (χ^2), entropy, information gain, and mutual information. The performances are evaluated using classification accuracy obtained from support vector machine, K nearest neighbors, and random forest classifiers on two publically available datasets viz., Stanford Twitter dataset and Ravikiran Janardhana dataset. In order to demonstrate the effectiveness of the proposed feature selection method, we conducted extensive experimentation by selecting different feature sets. The proposed feature selection method outperforms the existing feature selection methods in terms of classification accuracy on the Stanford Twitter dataset. Similarly, the proposed method performs competently equally in terms of classification accuracy compared to other feature selection methods in most of the feature subsets on Ravikiran Janardhana dataset.

Keywords: Feature selection, sentiment analysis, short text, classification.

2010 Mathematics Subject Classification: 68U15.

1 Introduction

The popularity of micro-blog applications in the recent decade generates enormous amount of short textual information. Millions of users make use of micro-blog sites to express their opinion or sentiment related to a product, topic, or events which take place in day to day life [39]. An opinion may be regarded as statements in which the opinion holder makes specific claim about a topic using certain sentiment [8, 24]. Many marketing companies use micro-blog textual information to identify sentiments related to the product or an event [10, 58]. The information retrieved from micro-blogs may involve at least two specific issues: firstly, use of formal languages, all in electronic word-of-mouth, which may lead to misspellings and use of slang words. Secondly, the limited characters which may tend to shortened words or sentences making analysis difficult. The detection and analysis of sentiments in short texts is an attractive topic, for many researchers and practitioners, to classify text into different polarities or classes.

A sentiment analysis is a process of automatically extracting opinions or emotions from text, especially in user-generated textual content. Sentiment analysis is considered as a classification task which classifies text into positive, negative, or neutral classes [4, 7, 16, 35, 55]. In order to create an automated system that performs an effective sentiment analysis, several researchers [23, 32–34, 50] came up with two main

***Corresponding author: H. M. Keerthi Kumar,** Department of Computer Science, JSS Research Foundation, JSS TI Campus, Campus Rd, Manasagangotri, Mysuru, Karnataka 570006, India, e-mail: hmkeerthikumar@gmail.com
B. S. Harish: Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, Karnataka, India

approaches: semantic orientation [3, 50] and machine learning method [33, 52]. Semantic orientation-based approach for sentiment analysis encompasses lexicon-based [47] and linguistic methods [46]. It has been claimed that lexicon-based and linguistic methods do not perform well on sentiment classification, due to the nature of an opinionated text, which requires more understanding of the text [48]. In addition to lexicon-based and linguistic methods, machine learning methods have been widely used for sentiment analysis [20]. In literature [6, 14, 18], machine learning-based approaches yield better predictive performance for sentiment analysis compared to lexicon-based methods. Generally, sentiment analysis based on machine learning algorithms can be performed using five steps viz., preprocessing, feature extraction and selection, representation, classification or clustering, and evaluation [17, 20]. Sentiment analysis on short texts needs to deal with high dimensionality of the features, due to low occurrence rate of feature across short texts. Most of the features are irrelevant and lead to poor performance of the classifier [36]. Therefore, selecting relevant features reduces the size of the feature space without sacrificing the performance of the sentiment classification.

In sentiment analysis, feature selection is a method to identify a subset of features to achieve various goals: firstly, to reduce computational cost, secondly, to avoid over fitting, and thirdly, to enhance the classification accuracy of the model [54]. Feature selection methods can be broadly divided into three categories, as filter methods, wrapper methods, and embedded methods [40]. The filter method assesses the optimal subset of features by looking only at the underlying properties of the data. The feature relevance scores are calculated, and low-scoring features are eliminated. The optimal subsets of features are presented to the classifier [42]. Wrapper method evaluates these subsets of features by detecting the possible interactions between features and learning model, i.e. wrapped around features and learning model to get an optimal subset of feature [25]. Embedded methods make use of both filter and wrapper method to select the optimal subset of features which increases the performance of the classifier. The most popular feature selection methods reported in the literature are chi-square (χ^2), entropy, information gain (IG), and mutual information (MI). Further, the selected features are used for the subsequent training of the machine learning classifiers.

The conventional feature selection methods consider the distribution of the short texts containing the feature between the classes. However, they do not take into account the frequency of the features within the classes. Hence, it is noted that a feature that is characteristic of a class must frequently appear in greater numbers in short texts belonging to the class than in other classes. This motivated us to propose a new, simple yet effective feature selection method. The proposed feature selection method considers class-wise features by computing the relevant features from each class. To determine the efficacy of proposed feature selection method, the proposed feature selection method is compared with conventional feature selection methods such as chi-square (χ^2) [45], entropy [44], IG [57], and MI [5]. The proposed method is evaluated using classification accuracy obtained from SVM, KNN, and RF classifiers on two publically available datasets: Stanford Twitter dataset [12] and Ravikiran Janardhana dataset [37].

The remainder of this paper is organized as follows: Section 2 reviews the related work on feature selection methods for sentiment analysis. Section 3 describes methodologies and proposed work with illustration. Section 4 contains experimental results and discussion. Section 5 concludes along with future work.

2 Related Work

Recently, micro-blogs data like tweets, Facebook posts, and reviews are growing at an unprecedented rate [15, 28]. The vast amount of user-generated short textual information has made micro-blogs the largest data source of public opinion. In micro-blogs, users make spelling mistakes and use slang words while expressing their views or opinions. Moreover, these short texts contain enormous amount of noisy data like url, punctuation, and special symbols that need to be preprocessed. The major challenges of sentiment analysis on micro-blogs are limited text, slang terms, high dimensionality, and sparsity. The curse of dimensionality and sparsity are a major concern in sentiment analysis where noisy, irrelevant features are present in feature space. In order to deal with these challenges, many researchers [13, 21, 22, 26, 27, 31, 43, 49, 51] explored the various machine learning approaches and concentrated their studies to the curse of dimensionality and used feature selection methods to reduce high dimensional feature space. Zhang et al. [56] proposed a feature selection method by adopting an attractive hidden topic analysis and entropy-based feature ranking. The method uses latent semantic analysis to find the latent structure of “topics” or “concepts” in a text corpus. The entropy-based feature selection method is used to rank the features related to the topic. The maximum entropy classifiers are used to evaluate the performance while curtailing the feature space significantly.

Zheng et al. [59] explored the effects of feature selection on sentiment analysis on Chinese online reviews. The N-char-grams and N-POS-grams are used to select potential sentimental features. The feature subsets are selected by using improved document frequency method, and feature weights are calculated by adopting Boolean weighting method. The chi-square test is carried out to test the significance of experimental results. The result suggests that low order N-char-grams can achieve a better performance than higher order N-char-grams when taking N-char-grams as features. Omar et al. [30] conducted a series of experimental comparisons on various feature selection methods for Arabic sentiment classification. The performance of various feature selection methods like IG, principal components analysis, Relief-F, Gini index, uncertainty, and chi-square feature selection methods were studied. The naive Bayes (NB), support vector machine (SVM) and K nearest neighbor (KNN) classifiers were used to classify Arabic documents into different polarities. The experimental result shows that the use of feature selection method increases the performance of the classifier. The SVM classifier performed better compared to other classifiers for all feature selection methods.

Various experimental comparisons were conducted on prominent feature extraction for English review analysis in Agarwal and Mittal [2]. The features were extracted using unigram, bi-gram, bi-tagged feature, and dependency parsing tree-based features. Further, IG and minimum redundancy maximum relevancy feature selection methods were used to eliminate the noisy and irrelevant features from the feature vector. SVM and multinomial NB classifiers were used to classify the review document into positive or negative class. The result showed that the multinomial NB performs better than SVM in terms of accuracy and execution time for binary sentiment classification. Wu et al. [53] proposed an improved text feature selection, based on text word frequency information. The method modifies the expected cross entropy algorithm using the following aspects: the frequency distribution within category and the frequency distribution among different categories. The experimental result shows that feature selection method based on occurrence of terms within different classes is essential in reducing feature space and in improving the performance of the classifier.

In literature, many researchers developed various feature selection methods for sentiment analysis. The existing feature selection methods consider the distribution of the short texts containing the feature between the classes. However, they do not take into account the frequency of the features within the classes. Hence, it is noted that a feature which is characteristic of a class must frequently appear in greater numbers in short texts belonging to the class than in other classes. The proposed method selects the frequently distributed features related to each class. This feature selection method is based on class-wise information to identify the relevant feature related to each class. The proposed feature selection method is evaluated using classification accuracy on three classifiers: SVM, KNN, and random forest (RF) classifiers.

3 Proposed Methodology

This section presents a detailed description of the methodology used for sentiment classification. Section 3.1 describes various preprocessing techniques used to eliminate less informative data from the dataset. Section 3.2 briefs text representation used in the proposed method. Section 3.3 gives a detailed description of the proposed feature selection method with an illustration. Finally, Section 3.4 briefs the classifiers used to classify sentiments into positive, negative, and neutral classes.

3.1 Preprocessing

Preprocessing involves the elimination of trivial or less informative data, which does not contribute to the sentiment classification. We used eight preprocessing techniques to process tweets, which are the following:

In tweets, user posted a **URL** along with text to provide supporting information about the text such as "http://bit.ly/IMXUM", which does not contribute to sentiment analysis. Hence, URL is replaced with white space.

Usually, tweets consists of **username (@)**, which implies or indicates the user. This username does not contribute much to the sentiment present in the tweets. Hence, we replace username with white space.

Hashtag (#) is associated with the particular topic and opinion expressed by the user in the tweets. We removed only the symbol "#", retaining the contents.

Negations play a vital role in sentiment classification; the co-occurrence of the negative word e.g. "not", "n't", etc., changes the orientation of text into different polarity. Hence, **negation handling** is used to expand short terms such as "don't", "can't", "n't", etc., terms to "do not", "cannot", "not", etc.

Usually, tweets contain exaggeration of terms such as "looovvvveee", and it is necessary to deal with these words to make them more formal. Hence, **characters normalization** is applied to replace consecutive characters, such as a character that appears more than three times to a single character.

Punctuation symbols such as ",", " ' ", "$", "?", "!", etc. do not contribute to the sentiment of tweets and are thus removed from the tweets. Finally, **stop-words** are eliminated and **stemming** is applied on each tweet.

3.2 Representation

The preprocessed short texts are represented in machine understandable forms. The preprocessed short texts are generally represented as vectors of terms using a bag of words [41] and n-gram (unigram and bigram) [38]. The work of [4] suggests that unigram with term frequency (*tf*) performs well on sentiment analysis for micro-blogging data. Hence, we have used unigram representation model, which is similar to Bag of Words model. Each word is considered as a term, and term frequency schema is used to calculate the frequency of terms appearing in each short text. The term weights are calculated by term frequency (tf_{t_i}) schema, i.e. tf_{t_i} = number of times term t_i appeared in a short text, where t_i represent the terms (features) present in the short text.

3.3 The Proposed Features Selection Method

In this section, we propose a novel feature selection method based on class-wise information. The class-wise feature selection method comprises three steps: firstly, the class-related short texts are grouped; secondly, the sum of the frequency of each feature corresponding to class are calculated. The obtained weights of feature values are sorted in descending order, and low weighted features are eliminated by fixing the threshold value. Here, threshold value is fixed empirically which indicates the number of features selected from each class. These processes are repeated for each class. Finally, the selected subsets of features from each class are combined to get overall features. These features are used for the subsequent training of the classifiers.

Let there be j number of classes and each class contains k number of short texts. The short texts are described by N dimensional term frequency vector (feature vector). The term document matrix, say S of size ($jk \times N$), is constructed such that each row represents a short text related to class C_j and each column represents a feature F, say $F = \{f_1, f_2, \dots, f_N\}$.

Firstly, we compute the sum of the frequency of each feature f_i corresponding to class C_j, i.e.

$$ClassTermFrequency(C_j, f_i) = \sum_{st=1}^{k} Frequency(S_{st}, f_i) \quad (1)$$

where $Frequency(S_{st}, f_i)$ is the frequency of occurrence of the features f_i in short text S_{st} and k is the number of short texts in the class C_j.

The size of the resultant $ClassTermFrequency(C_j, f_i)$ matrix will be $1 \times N$ for each class. Further, we sort the values of *ClassTermFrequency* in descending order to get the most frequent occurrences of terms within the classes. A subset N' features are selected by fixing threshold values. Here, threshold value is fixed based on empirical evaluation. The selected features are $F'_1 = \{f_1, f_2, \ldots, f_{N'}\}$, where $N' < N$ and F'_1 are features corresponding to a class. Similarly, we repeat the procedure for each class. Further, we apply union function to feature sets obtained from each class i.e.

$$F' = F'_1 \bigcup F'_2 \ldots \bigcup F'_j \tag{2}$$

The obtained F' features from the above computation are used for the subsequent training of the classifiers.

3.4 Illustration

In this section, a detailed illustration of individual steps involved in the proposed method is explained on the term document matrix S. Initially, short texts are preprocessed using various preprocessing techniques and represented using unigram representation model with term frequency (tf_{t_i}) schema. Table 1 presents an example of term document matrix say S for $k = 6$, $j = 2$, $t = 15$. Here, k denotes the number of short texts, j denotes the number of classes, and t denotes the number of terms as features.

In the first step, the class-related short texts are grouped to compute the relevant features which contribute towards classes. The short texts S_1, S_2, and S_3 represent the term document matrix related to class C_1. Similarly, short texts S_4, S_5, and S_6 present the term document matrix related to class C_2.

In the next step, *ClassTermFrequency* is computed for each class.

The *ClassTermFrequency* gives the sum of the frequency of features appearing in each class. Tables 2 and 3 give the results of the computation. The obtained matrix will be in the form $1 \times N$, which consists of N dimensional feature vector.

Further, we arrange the computed *ClassTermFrequency* in descending order based on the weight of features associated with the classes. Tables 4 and 5 show the results of the computations. The resultant matrix gives the highly relevant features that contribute to the classes.

In the next step, we select the threshold value for N' for each class. The threshold value is arrived at through multiple iterations of considering different values. We considered different values and arrived at the threshold value N'. $N' = 5$ is observed as the best value for the given example, where N' is less than N.

Table 1: Term Document Matrix (S).

Short text	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	
1	2	3	4	0	1	2	3	4	2	1	3	4	1	2	1	C_1
2	1	4	2	1	0	2	2	1	0	1	0	0	2	1	2	
3	1	1	0	2	3	3	1	0	1	0	0	0	0	0	0	
4	4	0	1	2	3	2	1	0	0	0	1	0	0	1	0	C_2
5	2	3	1	0	2	3	4	1	2	0	0	0	1	0	1	
6	1	2	3	1	4	1	0	1	1	0	0	1	1	0	1	

Table 2: Computation of *ClassTermFrequency* for Class 1.

Class	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}
1	4	8	6	3	4	7	6	5	3	2	3	4	3	3	3

Table 3: Computation of *ClassTermFrequency* for Class 2.

Class	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}
2	7	5	5	3	9	6	5	2	3	0	1	1	2	1	2

Table 4: Arranging *ClassTermFrequency* in Descending Order in Class 1.

Class	t_2	t_6	t_3	t_7	t_8	t_1	t_5	t_{12}	t_4	t_9	t_{11}	t_{13}	t_{14}	t_{15}	t_{10}
1	8	7	6	6	5	4	4	4	3	3	3	3	3	3	2

Table 5: Arranging *ClassTermFrequency* in Descending Order in Class 2.

Class	t_5	t_1	t_6	t_2	t_3	t_7	t_4	t_9	t_8	t_{13}	t_{15}	t_{11}	t_{12}	t_{14}	t_{10}
2	9	7	6	5	5	5	3	3	2	2	2	1	1	1	0

Table 6: Selecting $N' = 5$ for Class 1.

Class	t_2	t_6	t_3	t_7	t_8
1	8	7	6	6	5

Table 7: Selecting $N' = 5$ for Class 2.

Class	t_5	t_1	t_6	t_2	t_3
2	9	7	6	5	5

The resultant matrix will be the relevant feature vector for each class. Tables 6 and 7 give the N' feature vector where $N' < N$. The selected features for class 1 are $F'_1 = \{t_2,\ t_6,\ t_3,\ t_7,\ t_8\}$ and for class 2 are $F'_2 = \{t_5,\ t_1,\ t_6,\ t_2,\ t_3\}$.

Further, F' is composed of the union of the first N' selected feature vector for each class i.e. $F' = F'_1 \bigcup F'_2$. The selected features are $F' = \{t_1,\ t_2,\ t_3,\ t_5,\ t_6,\ t_7,\ t_8\}$.

Finally, the selected features F' are used for the subsequent training of the classifiers.

3.5 Classification

In order to evaluate the performance of the proposed feature selection method, we used three classifiers: SVM, KNN, and RF. SVM is a widely used classifier in sentiment classification tasks. It can effectively conduct classification tasks in higher-dimensional feature space [29]. On the other hand, the objective of KNN classifier is to classify based on majority vote of its neighbors, with the object being assigned to the class most common among its KNN. Here “K” indicates the number of neighbors taken into account in determining the class [1]. RF operates by constructing a multitude of decision trees at training time and outputting the class based on the decision of individual trees [9].

4 Experimental Evaluation

In this section, we present experimentation of the proposed method, and results are compared with existing feature selection methods.

4.1 Dataset Description

The experimentation was conducted on two publicly available datasets: Stanford Twitter Sentiment test dataset (Dataset 1) [12] and Ravikiran Janardhana dataset (Dataset 2) [37]. Dataset 1 contains 498 tweets that come with labels of 182 positive, 177 negative, and 139 neutral tweets. The total number of features obtained after preprocessing (as explained in Section 3.1) is 1586 features. Dataset 2 consists of 9666 positive, 9667 negative, and 2271 neutral tweets, which are combinations of [19] and [11] publicly available twitter message datasets. Here, we randomly choose 6000 tweets from Dataset 2 as overall data, where equal proportions of positive, negative, and neutral tweets are taken for experimentation. We obtained 10,349 features after preprocessing (as explained in Section 3.1) technique.

4.2 Experimental Setup

In this section, we compared the performance of the proposed feature selection method with chi-square (χ^2), entropy, IG, and MI feature selection methods. The experimentation is conducted under three splits of 50:50, 60:40, and 70:30 proportions of training and testing data. In the experiments, 10-fold cross-validation method is utilized. The evaluation of the feature selection methods is based on the classification accuracy obtained from SVM, KNN, and RF classifiers. The experiment was conducted using statistical computing toolkit R language version R-3.1.3. In this experiment, we have used linear kernel SVM classifier, which is considered as the basic form of SVM to classify the text corpus to different polarities or classes. In KNN, K value is fixed empirically as 3, which gives higher accuracy than any other values. RF produces multi-altitude decision trees at input phase, and the output is generated in the form of multiple decision trees. Here, the number of trees (ntree $=$ 100) is considered empirically, which gives higher accuracy as compared to other values.

4.2.1 Dataset 1 (Stanford Twitter dataset – 498 tweets)

Before performing the classification task, the short texts are preprocessed. The total number of features obtained after preprocessing was 1586 distinct features. The proposed feature selection method was applied by selecting threshold values as 100, 300, and 500 related to each class based on the empirical evaluation. The total number of features obtained for 100 is 220 features, for 300 is 678 features, and for 500 is 1154 features. Further increase in the threshold values achieved the original features i.e. 1586 features. Similarly, decrease in the threshold value led to a very small feature set which would not yield good results. Hence, we restricted the threshold values to be between 100 and 500. The obtained feature subsets from each threshold values are taken for comparison with chi-square (χ^2), entropy, IG, and MI feature selection methods. The experimental results are presented in Table 8.

In the first set of experiments (50:50 split), the classification accuracy obtained for the original 1586 features using SVM is 77.60%, 78.40% using KNN and 77.20% using RF. Table 8 presents the classification accuracy using the feature selection methods χ^2, entropy, IG, and MI and proposed feature selection method using SVM, KNN, and RF classifier on varying feature subsets. From the observations, it is noted that the RF classifier achieves maximum accuracy of 81.60% for 678 features compared to the other two classifiers.

In the second set of experiments (60:40 split), the proposed feature selection method using RF classifier exhibits the same classification accuracy of 83.50% for 220 and 678 features. However, in Table 8 the classification accuracy of RF classifier on the proposed feature selection method increases by 9.5%, 5.17%, 3.5%, and 2.84% for 220 features compared to χ^2, entropy, IG, and MI, respectively. On the other hand, for the classification accuracy of the proposed feature selection method using RF classifier for 678 features, there is an increase of 9%, 4.84%, 6.5%, and 5.5% as compared to χ^2, entropy, IG, and MI, respectively. Hence, RF classifier performs better for the second set of experimentation.

In the third set of experiments (70:30 split), the classification accuracy obtained for original features using SVM is 81.45%, 85.00% using KNN, and 83.50% using RF. From Table 8, we can observe that the proposed feature selection method achieves a maximum classification accuracy of 86.09% and 86.75% for 1154 features using SVM and KNN classifiers, respectively. The proposed feature selection method achieves better classification accuracy of 90% for 220 features using RF classifier compared to the other classifiers.

4.2.2 Dataset 2 (Ravikiran Janardhana dataset)

Initially, short texts were preprocessed using various preprocessing techniques and represented using the unigram model. The total number of features obtained after preprocessing are 10,349 distinct features. We

Table 8: Classification Accuracy on Dataset 1.

Training vs. testing	For all 1586 features			Features	Feature selection														
	SVM	KNN	RF		Chi-square			Entropy			Information gain			Mutual information			Proposed method		
					SVM	KNN	RF	SVM	KNN	RF	SVM	KNN	RF	SVM	KNN	RF	SVM	KNN	RF
50:50	77.60	78.40	77.20	220	72.00	72.80	70.80	69.33	66.40	76.53	71.73	69.60	79.20	71.73	68.53	78.93	76.40	79.20	78.40
				678	74.00	67.20	72.40	71.73	61.60	75.20	69.60	57.33	77.33	68.53	58.40	75.73	80.00	80.00	81.60
				1154	75.60	68.00	71.20	69.06	58.13	75.46	68.00	62.66	74.66	66.13	58.40	76.26	80.80	75.60	80.40
60:40	79.00	81.00	83.50	220	73.40	77.50	74.00	67.33	66.33	78.33	75.33	69.33	80.00	76.76	70.00	80.66	82.50	80.50	83.50
				678	76.00	79.00	74.50	69.66	60.66	78.66	68.66	57.66	77.00	67.33	58.33	78.00	83.00	82.50	83.50
				1154	80.50	71.00	81.00	66.66	58.33	78.00	67.33	63.66	77.66	65.66	58.66	76.33	82.00	80.00	83.00
70.30	81.45	85.00	83.50	220	74.83	81.45	75.49	72.00	65.77	78.33	74.66	70.22	80.88	73.77	70.66	80.00	78.80	84.10	90.00
				678	82.11	84.10	82.11	66.22	61.77	78.66	69.33	59.11	77.33	68.00	59.11	78.66	81.45	85.43	87.41
				1154	76.15	80.79	79.47	65.22	59.11	78.00	66.66	65.33	76.88	66.22	59.11	78.22	86.09	86.75	82.11

applied the proposed feature selection method by selecting the threshold values of 3500, 4000, 4500, 5000, and 5500 related to each class based on empirical evaluation. The total number of features obtained was 7200, 7922, 8649, 9359, and 10,084 features, respectively, for each of the respective threshold values. Further, by increasing the threshold values, we achieved the original features i.e. 10,349 features. Hence, we restricted the threshold value to 5500. Similarly, by decreasing the threshold value, we achieved very few feature sets which would not yield good results. Hence, we restricted the threshold value between 3500 and 5500. The obtained feature sets were evaluated using chi-square (χ^2), entropy, IG, and MI feature selection methods. The experimental results are tabulated in Table 9.

In the first set of experiments (50:50 split), the classification accuracy obtained for the original features using SVM is 74.30%, 62.90% using KNN, and 78.70% using RF. Table 9 presents the classification accuracy using SVM, KNN, and RF classifiers, respectively. From Table 9, it can be observed that the proposed feature selection method achieves 68.33% classification accuracy for 7200 features using KNN classifiers. On the other hand, RF classifier achieves maximum accuracy of 79.46% for 10,084 features with increase of 5.6%, 0.98%, 1.09%, 0.31% for chi-square (χ^2), entropy, IG, and MI feature selection methods, respectively. From the observations, it is noted that the RF classifier achieves better results when compared to the other two classifiers.

In the second set of experiments (60:40 split), the classification accuracy obtained for the original features using SVM is 74.70%, 65.00% using KNN, and 84.36% using RF. From Table 9, it is evident that the IG feature selection method achieves a maximum accuracy for 7200 features. However, from Table 9 we can observe that the proposed feature selection method achieves 84.69% for 10,084 features with an increase of 6.89% for χ^2, 0.47% for entropy, 0.19% for IG, and 0.14% for MI features selection method using RF classifier.

In the third set of experiments (70:30 split), Table 9 depicts that the IG feature selection method gives better result for 7200 features using SVM and RF classifier. However, the proposed feature selection method achieves a classification accuracy of 70.51% for 9359 features using KNN classifier with increase in 5.11%, 0.21%, 2.85%, and 2.92% for χ^2, entropy, IG, and MI features selection methods, respectively. The proposed feature selection method performs competently similarly in terms of classification accuracy to IG and MI feature selection methods in most of the feature subsets using RF classifier.

4.3 Discussion

It is evident from Table 8 that the proposed feature selection method performs better than the chi-square χ^2, entropy, IG, and MI feature selection methods using SVM, KNN, and RF classifiers on the Stanford Twitter Sentiment dataset. From Table 8, we can infer that, in terms of classification accuracy, RF performed better compared to the other classifiers. On the other hand, the proposed feature selection method was also experimented on in the Ravikiran Janardhana dataset. The proposed feature selection method considers the frequency of the features distributed within the class rather than the frequency distribution between the classes. The χ^2 score was calculated based on the term independent from the class. Thus, the proposed feature selection method performs better than the chi-square feature selection, on both datasets. Entropy measures the uncertainty of a distribution, which expresses the average amount of information contained in a text. In the proposed feature selection method, features corresponding to classes are considered to select the most relevant feature. Thus, the proposed feature selection method performs better using entropy feature selection method on both datasets.

On the other hand, the IG and MI scores are calculated based on the probabilities of terms or features occurrences in the classes. In IG, the scores are computed based on conditional probability of a class for a given term and entropy. IG considers presence or absence of the term or feature in a given input text. Dataset 1 consists of fewer numbers of presence or absence of features compared to Dataset 2. However, the proposed method purely depends on the frequency of the features distributed within the classes. Therefore, the proposed feature selection method on Dataset 2 performs reasonably good compared to IG in most of the feature subsets than Dataset 1. Similarly, MI is strongly influenced by the marginal probabilities of the features where it measures the dependencies between random terms or features. It can be observed from Table 9 that Dataset 2 consists of higher number of features than Dataset 1. The proposed method depends on

Table 9: Classification Accuracy on Dataset 2.

Training vs. testing	For all 10,349 features			Features	Feature selection														
	SVM	KNN	RF		Chi-square			Entropy			Information gain			Mutual information			Proposed method		
					SVM	KNN	RF	SVM	KNN	RF	SVM	KNN	RF	SVM	KNN	RF	SVM	KNN	RF
50:50	74.30	62.90	78.70	7200	68.90	63.70	77.95	74.10	65.00	78.55	70.51	58.84	78.97	76.57	65.42	79.35	74.35	68.33	78.53
				7922	73.40	60.00	77.82	74.00	64.90	78.17	70.62	67.28	79.13	76.60	65.33	79.77	75.00	64.95	78.86
				8649	71.60	60.90	77.86	74.60	64.70	78.64	71.37	67.42	79.00	76.35	64.60	80.26	74.88	68.30	78.71
				9359	74.00	62.40	72.60	74.40	64.60	78.43	70.22	64.22	79.40	76.00	64.06	79.71	74.46	65.00	78.93
				10084	74.00	62.80	73.86	74.20	63.10	78.48	75.64	63.86	78.37	76.02	64.04	79.15	74.90	64.70	79.46
60:40	74.70	65.00	84.36	7200	74.50	64.20	78.00	74.60	66.10	84.25	78.47	70.11	84.69	76.91	66.77	84.52	75.40	63.90	84.63
				7922	74.80	60.90	77.52	75.10	66.00	84.36	77.16	69.83	84.66	76.33	66.66	85.00	74.80	66.10	84.61
				8649	74.50	62.30	78.19	75.20	65.70	84.63	77.08	69.72	84.33	77.02	66.55	85.08	74.83	65.70	84.55
				9359	74.60	64.20	73.69	75.30	65.60	84.18	76.08	66.11	84.44	76.02	66.08	84.75	74.80	65.60	84.72
				10084	72.60	64.90	77.80	75.20	65.00	84.22	75.50	65.44	84.50	75.50	65.61	84.55	75.02	65.00	84.69
70:30	74.90	66.30	84.63	7200	76.10	64.90	78.37	75.90	71.10	84.37	78.03	68.44	85.51	77.37	71.77	58.29	73.81	68.14	85.44
				7922	74.50	61.10	78.55	75.40	71.10	85.11	77.70	70.85	85.14	77.74	71.66	85.66	75.92	71.30	85.44
				8649	74.90	63.00	74.40	75.20	70.80	84.40	76.70	70.51	85.22	77.18	71.07	85.07	75.50	71.00	85.14
				9359	75.90	65.40	74.40	75.70	70.30	84.96	77.25	67.66	85.29	77.51	67.59	85.14	75.20	70.51	84.92
				10084	74.80	66.20	77.22	75.60	70.00	84.07	76.55	66.74	84.88	76.07	67.00	85.37	75.51	70.03	84.81

the frequency of features to select discriminative features rather than probabilities of two random features. Therefore, the proposed feature selection method performs competitively better compared to MI in most of the feature subsets on Dataset 2. The overall results show that the proposed feature selection method outperforms other feature selection methods in terms of classification accuracy on Dataset 1. On the other hand, the proposed feature selection method on Dataset 2 significantly outperforms chi-square and entropy feature selection methods. In case of IG and MI feature selection methods, competitive result can be found in most of the feature subsets in terms of classification accuracy.

5 Conclusion and Future Work

Sentiment analysis on short text is a recent and active area of research. In short text, there are many challenges that need to be addressed i.e. use of formal language, misspellings, and shortened form of words, which leads to high dimensionality and sparsity. To deal with these challenges, in this paper, we proposed a novel, simple, and yet effective feature selection method based on frequently distributed features related to each class. The experimental results of the proposed feature selection method are compared with chi-square (χ^2), entropy, IG, and MI feature selection methods using SVM, KNN, and RF classifiers, on two publically available datasets. The experimental result shows that the proposed feature selection method outperforms other feature selection methods in terms of classification accuracy on Dataset 1. On the other hand, the proposed feature selection method performs competently similarly in terms of classification accuracy to IG and MI feature selection methods in most of the feature subsets on Dataset 2.

In future, we would like to amalgamate (a) the statistical methods for calculating threshold values and (b) the n-gram representation (bigrams and trigrams) on the proposed feature selection using different classifiers which could further enhance the classification performance.

Acknowledgments: H. M. Keerthi Kumar has been financially supported by UGC under Rajiv Gandhi National Fellowship (RGNF) Letter no. F1-17.1/2016-17/RGNF-2015-17-SC-KAR-6370/(SA-III Website), JSSRF (University of Mysore), Karnataka, India.

Bibliography

[1] D. A. Adeniyi, Z. Wei and Y. Yongquan, Automated web usage data mining and recommendation system using K-nearest neighbor (KNN) classification method, *Appl. Comput. Inform.* **12** (2016), 90–108.

[2] B. Agarwal and N. Mittal, Prominent feature extraction for review analysis: an empirical study, *J. Exp. Theor. Artif. Intell.* **28** (2016), 485–498.

[3] B. Agarwal and N. Mittal, Semantic orientation-based approach for sentiment analysis, in: *Prominent Feature Extraction for Sentiment Analysis*, pp. 77–88, Springer, Cham, Switzerland, 2016.

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, Sentiment analysis of twitter data, in: *Proceedings of the Workshop on Languages in Social Media*, pp. 30–38, Association for Computational Linguistics, Portland, Oregon, 2011.

[5] D. Agnihotri, K. Verma and P. Tripathi, Variable Global Feature Selection Scheme for automatic classification of text documents, *Expert Syst. Appl.* **81** (2017), 268–281.

[6] A. Al-Saffar, S. Awang, H. Tao, N. Omar, W. Al-Saiagh and M. Al-bared, Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm, *PLoS One* **13** (2018), e0194852.

[7] R. K. Amplayo and M. Song, An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews, *Data Knowl. Eng.* **110** (2017), 54–67.

[8] M. R. Bouadjenek, H. Hacid and M. Bouzeghoub, Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms, *Inform. Syst.* **56** (2016), 1–18.

[9] A. Bouaziz, C. Dartigues-Pallez, C. da Costa Pereira, F. Precioso and P. Lloret, Short text classification using semantic random forest, in: *International Conference on Data Warehousing and Knowledge Discovery*, pp. 288–299, Springer, Cham, Switzerland, 2014.

[10] M. S. Checkley, D. Añón Higón and H. Alles, The hasty wisdom of the mob: how market sentiment predicts stock market behavior, *Expert Syst. Appl.* **77** (2017), 256–263.

[11] Corpus, *Sanders-Twitter Sentiment*, http://www.sananalytics.com/lab/twitter-sentiment/sanders-twitter-0.2.zip. Accessed 10 October, 2017.

[12] [Dataset], *Sentiment140 corpus*, http://help.sentiment140.com/for-students/. Accessed 10 November, 2018.

[13] M. del Pilar Salas-Zarate, M. A. Paredes-Valverde, J. Limon, D. A. Tlapa and Y. A. Báez, Sentiment classification of spanish reviews: an approach based on feature selection and machine learning methods, *J. Univers. Comput. Sci.* **22** (2016), 691–708.

[14] M. D. Devika, C. Sunitha and A. Ganesh, Sentiment analysis: a comparative study on different approaches, *Procedia Comput. Sci.* **87** (2016), 44–49.

[15] C. Francalanci and A. Hussain, Influence-based Twitter browsing with NavigTweet, *Inform. Syst.* **64** (2017), 119–131.

[16] G. Ganu, Y. Kakodkar and A. Marian, Improving the quality of predictions using textual information in online user reviews, *Inform. Syst.* **38** (2013), 1–15.

[17] G. Gautam and D. Yadav, Sentiment analysis of twitter data using machine learning approaches and semantic analysis, in: *Contemporary Computing (IC3), 2014 Seventh International Conference on*, pp. 437–442, IEEE, Noida, India, 2014.

[18] G. Gezici, B. Yankoğlu, D. Tapucu and Y. Saygn, New features for sentiment analysis: do sentences matter? in: *CEUR Workshop Proceedings*, Bristol, UK, 2012.

[19] A. Go, R. Bhayani and L. Huang, Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford* **1** (2009), 12.

[20] E. Haddi, X. Liu and Y. Shi, The role of text pre-processing in sentiment analysis, *Procedia Comput. Sci.* **17** (2013), 26–32.

[21] B. S. Harish and M. B. Revanasiddappa, A comprehensive survey on various feature selection methods to categorize text documents, *Int. J. Comput. Appl.* **164** (2017), 1–7.

[22] C. Huang, J. Zhu, Y. Liang, M. Yang, G. Pui, C. Fung and J. Luo, An efficient automatic multiple objectives optimization feature selection strategy for internet text classification, *Int. J. Mach. Learn. Cyb.* **9** (2018), 1–13.

[23] C. Hung, Word of mouth quality classification based on contextual sentiment lexicons, *Inform. Process. Manag.* **53** (2017), 751–763.

[24] S.-M. Kim and E. Hovy, Determining the sentiment of opinions, in: *Proceedings of the 20th International Conference on Computational Linguistics*, p. 1367, Association for Computational Linguistics, Geneva, Switzerland, 2004.

[25] R. Kohavi and G. H. John, Wrappers for feature subset selection, *Artif. Intell.* **97** (1997), 273–324.

[26] S. Kübler, C. Liu and Z. A. Sayyed, To use or not to use: feature selection for sentiment analysis of highly imbalanced data, *Nat. Lang. Eng.* **24** (2018), 3–37.

[27] A. Kumar and R. Khorwal, Firefly algorithm for feature selection in sentiment analysis, in: *Computational Intelligence in Data Mining*, pp. 693–703, Springer, Singapore, 2017.

[28] B. Li, K. C. C. Chan, C. Ou and S. Ruifeng, Discovering public sentiment in social media for predicting stock movement of publicly listed companies, *Inform. Syst.* **69** (2017), 81–92.

[29] H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, *IEEE Trans. Knowl. Data Eng.* **17** (2005), 491–502.

[30] N. Omar, M. Albared, T. Al-Moslmi and A. Al-Shabi, A comparative study of feature selection and machine learning algorithms for Arabic sentiment classification, in: *Asia Information Retrieval Symposium*, pp. 429–443, Springer, Charm, Singapore, 2014.

[31] A. Onan and S. Korukoğlu, A feature selection model based on genetic rank aggregation for text sentiment classification, *J. Inf. Sci.* **43** (2017), 25–38.

[32] A. C. Pandey, D. S. Rajpoot and M. Saraswat, Twitter sentiment analysis using hybrid cuckoo search method, *Inform. Process. Manag.* **53** (2017), 764–779.

[33] B. Pang, L. Lee and S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, Philadelphia, 2002.

[34] I. Penalver-Martinez, F. Garcia-Sanchez, R. Valencia-Garcia, M. A. Rodriguez-Garcia, V. Moreno, A. Fraga and J. L. Sanchez-Cervantes, Feature-based opinion mining through ontologies, *Expert Syst. Appl.* **41** (2014), 5995–6008.

[35] D.-H. Pham and A.-C. Le, Learning multiple layers of knowledge representation for aspect based sentiment analysis, *Data Knowl. Eng.* **114** (2017), 26–39.

[36] R. H. W. Pinheiro, G. D. C. Cavalcanti, R. F. Correa and T. I. Ren, A global-ranking local feature selection method for text categorization, *Expert Syst. Appl.* **39** (2012), 12851–12857.

[37] J. Ravikiran, *Twitter sentiment analysis and opinion mining*, Data Mining Project Report, 2010.

[38] Y. Ren, R. Wang and D. Ji, A topic-enhanced word embedding for Twitter sentiment classification, *Inform. Sci.* **369** (2016), 188–198.

[39] F. Riquelme and P. González-Cantergiani, Measuring user influence on Twitter: a survey, *Inform. Process. Manag.* **52** (2016), 949–975.

[40] Y. Saeys, I. Inza and P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics* **23** (2007), 2507–2517.

[41] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Inform. Process. Manag.* **24** (1988), 513–523.

[42] N. Sánchez-Maroño, A. Alonso-Betanzos and M. Tombilla-Sanromán, Filter methods for feature selection – a comparative study, in: *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, 178–187, 2007.

[43] R. Shahid, S. T. Javed and K. Zafar, Feature selection based classification of sentiment analysis using Biogeography optimization algorithm, in: *Innovations in Electrical Engineering and Computational Technologies (ICIEECT), 2017 International Conference on*, pp. 1–5, IEEE, Karachi, Pakistan, 2017.

[44] C. E. Shannon, A mathematical theory of communication, *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **5** (2001), 3–55.

[45] F. Song, S. Liu and J. Yang, A comparative study on text representation schemes in text categorization, *Pattern Anal. Appl.* **8** (2005), 199–209.

[46] M. Taboada, Sentiment analysis: an overview from linguistics, *Annu. Rev. Linguist.* **2** (2016), 325–347.

[47] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, Lexicon-based methods for sentiment analysis, *Comput. Linguist.* **37** (2011), 267–307.

[48] M. Thelwall, K. Buckley and G. Paltoglou, Sentiment in Twitter events, *J. Assoc. Inform. Sci. Technol.* **62** (2011), 406–418.

[49] A. Tommasel and D. Godoy, A Social-aware online short-text feature selection technique for social media, *Inform. Fusion* **40** (2018), 1–17.

[50] P. D. Turney and M. L. Littman, Measuring praise and criticism: inference of semantic orientation from association, *ACM Trans. Inform. Syst. (TOIS)* **21** (2003), 315–346.

[51] A. K. Uysal and Y. L. Murphey, Sentiment classification: feature selection based approaches versus deep learning, in: *Computer and Information Technology (CIT), 2017 IEEE International Conference on*, pp. 23–30, IEEE, Helsinki, Finland, 2017.

[52] D. Vilares, M. A. Alonso and C. Gómez-Rodrguez, Supervised sentiment analysis in multilingual environments, *Inform. Process. Manag.* **53** (2017), 595–607.

[53] G. Wu, L. Wang, N. Zhao and H. Lin, Improved expected cross entropy method for text feature selection, in: *Computer Science and Mechanical Automation (CSMA), 2015 International Conference on*, pp. 49–54, IEEE, Hangzhou, China, 2015.

[54] A. Yousefpour, R. Ibrahim and H. N. Abdel Hamed, Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis, *Expert Syst. Appl.* **75** (2017), 80–93.

[55] N. Zainuddin and A. Selamat, Sentiment analysis using support vector machine, in: *Computer, Communications, and Control Technology (I4CT), 2014 International Conference on*, pp. 333–337, IEEE, Langkawi, Malaysia, 2014.

[56] Z. Zhang, X.-H. Phan and S. Horiguchi, An efficient feature selection using hidden topic in text categorization, in: *Advanced Information Networking and Applications-Workshops, 2008. AINAW 2008. 22nd International Conference on*, pp. 1223–1228, IEEE, Okinawa, Japan, 2008.

[57] D. M. Zhang, S. Li, C. Zhu, X. Niu and L. Song, A comparison study of multi-class sentiment classification for Chinese reviews, in: *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, 5, pp. 2433–2436, IEEE, Yantai, China, 2010.

[58] B. Zhao, Z. Zhang, W. Qian and A. Zhou, Identification of collective viewpoints on microblogs, *Data Knowl. Eng.* **87** (2013), 374–393.

[59] L. Zheng, H. Wang and S. Gao, Sentimental feature selection for sentiment analysis of Chinese online reviews, *Int. J. Mach. Learn. Cyb.* **9** (2015), 1–10.